A High-Speed Photonic Tensor Accelerator

Alireza Fardoost
CREOL, The College of Optics and Photonics, University of Central Florida
Orlando, FL, USA
a.fardoost@knights.ucf.edu

Fatemeh Ghaedi Vanani
CREOL, The College of Optics and Photonics, University of Central Florida
Orlando, FL, USA
f.ghaedi@knights.ucf.edu

Zheyuan Zhu
CREOL, The College of Optics and Photonics, University of Central Florida
Orlando, FL, USA
zyzhu@knights.ucf.edu

Christopher Doerr
Aloe Semiconductor, Inc. 1715 Highway 35, Suite 303
Middletwon, NJ, USA
crdoerr@ieee.org

Shuo Pang
CREOL, The College of Optics and Photonics, University of Central Florida
Orlando, FL, USA
pang@ucf.edu

Guifang Li
CREOL, The College of Optics and Photonics, University of Central Florida
Orlando, FL, USA
li@ucf.edu

2022 IEEE Photonics Conference (IPC) | 978-1-6654-3487-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/IPC53466.2022.9975540

***Abstract*— We propose a coherent multi-dimensional (wavelength, spatial mode, polarization, etc.) photonic tensor accelerator capable of matrix-vector, matrix-matrix, and batch matrix multiplications in a single clock cycle. A proof-of-concept 2x2 matrix-matrix multiplication at 25GBd with 4.67 bit precision was experimentally demonstrated.**

Keywords—Photonic Accelerator, Matrix Multiplication, Optics for AI, Analog Computing

I. Introduction

Artificial intelligence (AI) advanced rapidly owing to innovations in electronic hardware accelerators that propelled computing performances beyond the scaling limits of Moore's law. The main approach of the current technologies including graphic processing units (GPUs) and tensor processing units (TPUs) has been to design parallel computing structures along with optimized memory organization. However, the cost will eventually limit the parallelization, and physics will limit the chip efficiency. Therefore, it is necessary to develop new technologies far beyond today's capabilities to support the astonishing growth of AI computation. Passive linear optical circuits have the potentials to greatly reduce both computing and data-movement energy consumption. Low-power optical-to-electrical conversion has also recently been studied and shown promising results [1]. Accordingly, the field of optical artificial neural networks is experiencing a resurgence [2-4]. Here, for the first time, we combined wavelength-division multiplexing (WDM) and mode-division multiplexing (MDM) to encode matrices of unprecedented sizes. The proposed method of multidimensional encoding and coherent mixing supports vector, matrix, and tensor processing within a single clock cycle. Since reliable communication technologies can be deployed in modulation, multiplexing, and detection, the expected computation speed can go up to 10s of GHz and the energy efficiency can be optimized. We envision our photonics platform to be the building block for AI applications, including but not limited to, deep neural networks, real-time image processing, and dynamic control systems.

II. Photonic Tensor Accelerator (PTA)

Matrix multiplication consists of multiply and accumulate (MAC) operations on all the elements. The speed of a hardware accelerator is directly related to how many parallel MACs can be performed. In the proposed PTA, scalar multiplication is performed via interference and coherent detection. The weight $\left(E_w = w.\exp\left(j\omega_0 t\right)\right)$ and input $\left(E_x = x.\exp\left(j\omega_0 t\right)\right)$ electric fields are combined on a balanced photodetector (BPD) and the output photocurrent would be their scalar multiplication $w \times x$.

To extend the scalar multiplication to inner product of two vectors, the weight and input vector elements can be mapped to different wavelengths and, as a result, the BPD output will be $\sum_{n=1}^{N} w_n \times x_n$ where N is the length of the vectors. Analogous to the orthogonality of wavelength modes in time, spatial modes are orthogonal in space. Therefore, a similar mapping to spatial modes will lead to the same inner product of the vectors.

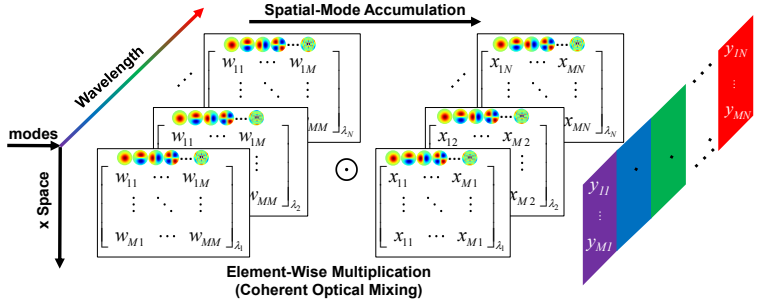

Fig. 1. Photonic Tensor Accelerator (PTA). Schematic mapping of the matrix elements on wavelengths, spatial modes, and space to implement matrix-matrix multiplication in a single clock cycle.

Combining WDM and MDM with parallelization in space will result in the PTA shown schematically in Fig. 1 capable of matrix-matrix multiplication in one clock cycle. As shown here, the W matrix is projected on the (mode, space) dimensions and duplicated in the wavelength dimension. Similarly, the X matrix is projected on the (mode, wavelength) dimensions and duplicated in space. The output matrix elements are mapped to different wavelengths and spatial locations.

Funding: This work was supported in part by the Office of Naval Research under contract N00014-20-1-2441, NSF under grant number ECCS-1932858, and the Army Research Office under contract W911NF2110321.

III. Matrix-Matrix Multiplicaiton Demonstration

The free-space implementation of a 2×2 matrix-matrix multiplier is illustrated in Fig. 2(a). Two CW lasers at $\lambda_1 = 1545nm$ and $\lambda_2 = 1555nm$ were modulated with a 25Gbps PRBS (2^{15}-1) using an EO modulator. To alleviate the requirements for a large number of high-speed equipment, we use only two modulators and appropriate delays so that all eight matrix elements are decorrelated with each other. A fiber delay is added between the two inputs of the all-fiber mode-selective photonic lantern (PL). Consequently, the output of the PL in two wavelengths and two spatial modes consists of four matrix elements of X ($x_{11}(\lambda_1, LP_{01})$, $x_{12}(\lambda_1, LP_{11a})$, $x_{21}(\lambda_2, \mathrm{LP}_{01})$, and $x_{22}(\lambda_2, LP_{11a})$). The interference and detection are performed in free space where the weight matrix W is a delayed and thus decorrelated version of the input matrix X. After passing through the WDM filters and balanced detection, the output matrix Y elements are obtained in one clock cycle.

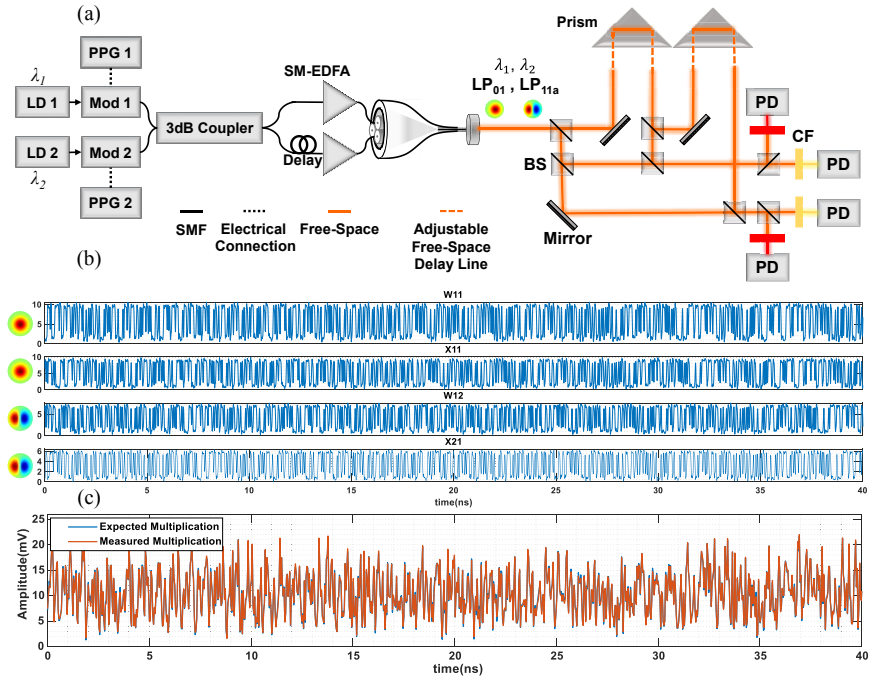

Fig. 2. Experiment setup and results for PTA in free space. (a) Schematic of the experimental setup for the matrix-matrix multiplication demonstration. LD: Laser Diode, PPG: Pulse Pattern Generator, BS: Beam Splitter, CF: Color Filter, PD: Photodetector. (b) Elements of the X and W matrices (c) measured multiplication results ($\tilde{Y}_{11}$ in red), in comparison with the expected result (Y_{T11} in blue).

Measured intensity waveforms for X and W, and one element $(\tilde{Y}_{11})$ of the matrix multiplication $(\tilde{Y})$ are shown for 1000 symbols $(40ns)$ in duration in Fig. 2(b) and (c), respectively. Using the measured X and W, their expected (ground truth) product (Y_T) can also be computed. The signal-to-noise ratio (SNR) of each output matrix element can be defined as $SNR = \left(V_{rms} / error_{rms} \right)^2$ where $error = Y_T - \tilde{Y}$ and V_{rms} is the root mean square (RMS) voltage of the measured signal. The SNR is found to be 21-22dB for different Y elements. Additionally, a more generalized parameter, normalized mean square error (NMSE), which is related to SNRs of all matrix elements, can be defined to characterize the accuracy of the entire matrix $\tilde{Y}$ [5]. To obtain the NMSE, a matrix is first vectorized and its L^2 norm is defined as $\left\| \vec{Y} \right\|^2 = \vec{Y} \cdot \vec{Y}^T$. The NMSE can then be calculated as $\left\| \vec{Y}_T - \vec{\tilde{Y}} \right\|^2 \Big/ \left\| \vec{Y}_T \right\|^2$.

The maximum number of detectable levels is equal to $1/\sqrt{NMSE}$. Finding NMSE for more than 10000 different symbols of the signal in time, the overall bit precision for the matrix multiplication can be denoted as $-\log_2 \left\langle \sqrt{NMSE} \right\rangle$ where $\langle . \rangle$ is the ensemble average over NMSEs. The experimental results show 25 detectable levels and a bit precision of 4.67 bits.

IV. Conclusion

The key innovation for the photonic tensor accelerator (PTA) lies in exploiting all dimensions of light, each containing many degrees of freedom. Because accumulation is multi-dimensional, the scalability of the proposed PTAs is multiplicative since these dimensions are orthogonal. The PTA also takes advantage of mature and reliable communication technologies.

References

1. D. A. Miller, Journal of Lightwave Technology **35**, 346-396 (2017).
2. X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, Science **361**, 1004-1008 (2018).
3. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, and D. Englund, Nature Photonics **11**, 441-446 (2017).
4. R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, Physical Review X **9**, 021032 (2019).
5. D. M. Allen, Technometrics **13**, 469-475 (1971).